Synthesizing State-of-the-Art Structure Predictions from Soup of Co-folding Models

Hyosoon Jang, **Taewon Kim**, **Sungsoo Ahn**

KAIST

Abstract. Co-folding models have advanced rapidly, yet no single model consistently performs best across all biomolecular complexes. This raises the question of whether independently trained co-folding models encode complementary information that can be transferred across co-folding models. We introduce SoupFold, which improves co-folding predictions by learning simple mappings between the representation spaces of co-folding models. At inference time, SoupFold transfers and incorporates representations from other co-folding models to update the representation used for structure prediction. Importantly, this does not re-train the co-folding models. We evaluate SoupFold on protein-protein and protein-ligand prediction tasks of FoldBench using AlphaFold3, Protenix, ESMFold2, and OpenDDE. By combining their representations, SoupFold achieves state-of-the-art performance on both protein-protein and protein-ligand structure prediction, showing that independently trained co-folding models encode complementary information that can be effectively transferred across models.

Correspondence: {hyosoon.jang,sungsoo.ahn}@kaist.ac.kr

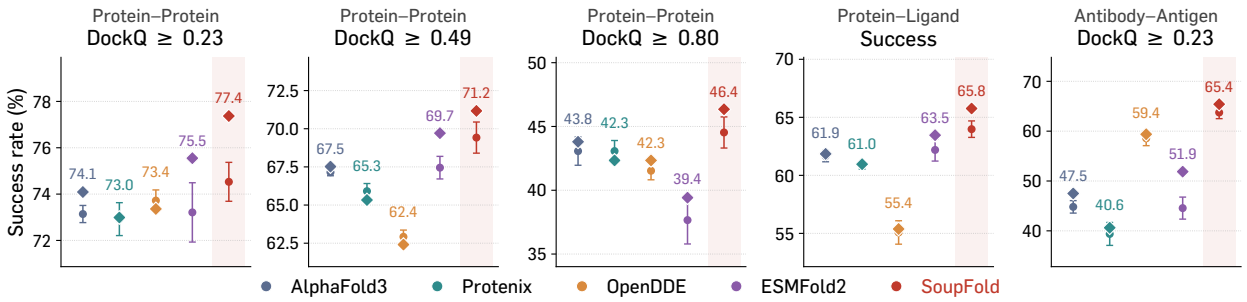

Figure 1 Co-folding structure prediction on FoldBench. For each model, ○ shows the mean ±1 s.d. of the success rate across five samples, and ◇ shows the best-of-five result selected by confidence. For antibody-antigen, we use 20 predictions per target to reduce the high variance of ESMFold2 when reproducing the reported results. Our SoupFold achieves the highest success rates.

1 Introduction

Co-folding models have advanced rapidly in recent years. Since AlphaFold3 (Abramson et al., 2024) extended structure prediction from single proteins to general biomolecular complexes, several models have followed, including Boltz (Passaro et al., 2025; Wohlwend et al., 2024), Chai (Chai Discovery et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), OpenFold3 (The OpenFold3 Team, 2026), ESMFold2(Candido et al., 2026), and OpenDDE (Aureka AI OpenDDE project, 2026). These models share a similar architecture: a Pairformer-based trunk constructs pairwise representations of a complex, which are then used by a diffusion module to generate its structure. They mainly differ in their training recipes including training data, objectives, and hyperparameters.

Despite rapid progress, no single co-folding model consistently performs best across all prediction targets. As illustrated in Figure 2, their relative performance varies across complexes. For example, ESMFold2 successfully predicts the left complex (7QPR A/B) but fails on the right (8Q6Q A/B), whereas other models

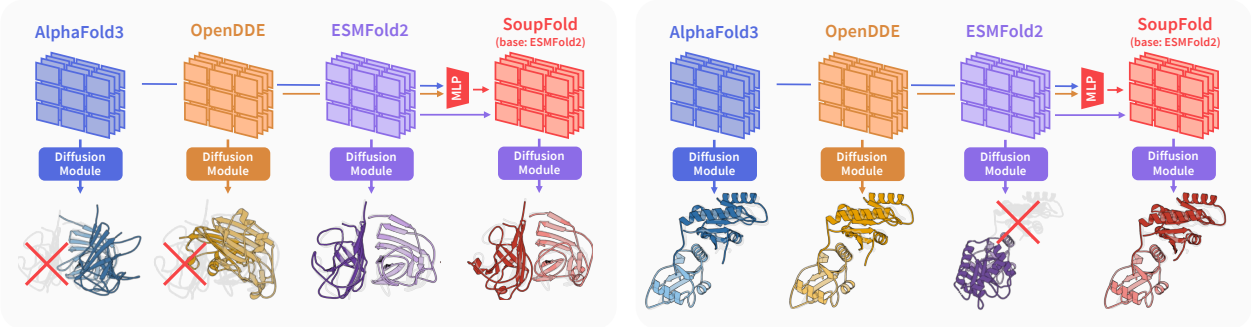

Figure 2 Examples of SoupFold on protein-protein prediction. At inference time, representations from multiple co-folding models are transferred into the representation space of a base model, e.g., ESMFold2, using lightweight MLP-based transfer networks. The transferred representations are then incorporated into the base representation before the base model's diffusion module generates the final structure. While individual co-folding models fail on either of the two complexes, SoupFold aggregates their representations and successfully predicts both.

show the opposite behavior. This complementary behavior suggests that different models may capture different structural patterns.

Here, we note that such differences may already be encoded in their internal representations. Since diffusion modules generate structures conditioned on pairwise representations, complementary predictions may arise from complementary information represented before structure generation. We thus ask whether transferring this information across models could improve the prediction. That is,

> *Do co-folding models encode complementary information in their representations, and can transferring this information improve another model?*

This question is closely related to representation distillation, where a student is trained to match the representation of a teacher (Ahn et al., 2019; Hinton et al., 2015). Recent work has applied this principle to generative models, showing that aligning internal representations with pretrained models improves generation for images (Yu et al., 2025) and protein structures (Kim et al., 2026). While prior work performs such transfer during training, we apply the same principle at inference time between state-of-the-art co-folding models.

Contribution. We show that co-folding models can significantly improve performance using information transferred from other independently trained co-folding models, even through a simple representation-level approach. As illustrated in Figure 2, we propose SoupFold, which transfers representations from teacher co-folding models, such as AlphaFold3 and OpenDDE, into the representation space of another co-folding model, such as ESMFold2. At inference time, SoupFold uses these transferred representations to update the base representation before structure generation. To our knowledge, this is the first work to combine independently trained co-folding models in representation space without retraining the co-folding models themselves.

We evaluate SoupFold on FoldBench using four co-folding models: AlphaFold3, Protenix-v1, ESMFold2, and OpenDDE-preview. By combining their representations at inference time, SoupFold achieves state-of-the-art performance on protein-protein and protein-ligand structure prediction tasks in FoldBench. We further show that SoupFold can recover predictions for targets where all four considered co-folding models fail.

2 Results

We evaluate whether complementary information from multiple co-folding models can be exploited to improve structure prediction. We consider four models with different architectures and training procedures: AlphaFold3 (Abramson et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), ESMFold2 (Candido et al., 2026), and OpenDDE-preview (Aureka AI OpenDDE project, 2026). We evaluate whether representations from the other models can improve each model in turn.

Table 1 Co-folding performance on FoldBench. Best-of-5 (confidence-selected) success rate; per-sample mean ± s.d. in parentheses. Protein-protein (274 interfaces): DockQ ≥ 0.23/0.49/0.80. Protein-ligand (517 interfaces, common subset): LRMSD < 2Å and lDDT-PLI > 0.8. All-Model-Best* is selected by maximum confidence over 20 samples generated by four co-folding models.

	Protein-Protein			Protein-Ligand
Model	Acceptable (≥ 0.23)	Medium (≥ 0.49)	High (≥ 0.80)	Success
All-Model-Best*	75.91	70.07	44.89	63.85
AlphaFold3	74.09 (73.14 ± 0.37)	67.52 (67.15 ± 0.23)	43.80 (43.07 ± 1.11)	61.85 (61.66 ± 0.49)
Protenix	72.99 (72.92 ± 0.71)	65.33 (65.91 ± 0.50)	42.34 (43.07 ± 0.83)	60.96 (60.85 ± 0.31)
+SoupFold	**75.91** (75.55 ± 1.13)	**70.07** (69.64 ± 0.54)	**44.53** (44.09 ± 0.54)	**63.44** (62.71 ± 0.59)
OpenDDE	73.36 (73.72 ± 0.46)	62.41 (62.92 ± 0.44)	42.34 (41.53 ± 0.71)	55.38 (55.08 ± 1.01)
+SoupFold	**79.20** (77.66 ± 1.04)	**68.25** (67.66 ± 0.50)	**44.53** (43.21 ± 1.23)	**63.44** (64.29 ± 0.79)
ESMFold2	75.55 (73.21 ± 1.28)	69.71 (67.45 ± 0.74)	39.42 (37.66 ± 1.87)	63.46 (62.19 ± 0.96)
+SoupFold	**77.37** (74.53 ± 0.84)	**71.17** (69.42 ± 1.02)	**46.35** (44.53 ± 1.22)	**65.76** (63.98 ± 0.71)

Benchmark. We mainly use protein-ligand and protein-protein complexes from FoldBench (Xu et al., 2026). For protein-ligand complexes, we report the success rate under LRMSD < 2 Å and lDDT-PLI > 0.8. For protein-protein complexes, we report the fraction of interfaces meeting the acceptable, medium, and high-quality thresholds of 0.23, 0.49, and 0.80, respectively. We evaluate 274 protein-protein interfaces and 517 protein-ligand interfaces that were successfully processed by all four models. We fix the recycling steps to 10 and use the default settings for all other configurations. We also consider antibody-antigen tasks in Section 5.

2.1 SoupFold Improves Structure Prediction Using Complementary Representations from Multiple Co-folding Models

We first ask whether representations from different co-folding models can be combined to improve structure prediction. SoupFold designates one model as the base model, e.g., ESMFold2, for structure generation and treats the remaining co-folding models, e.g., AlphaFold3, Protenix, OpenDDE, as teachers. We learn lightweight mappings between their pair representation spaces while keeping all co-folding models fixed. At inference time, teacher representations are mapped into the base model's space and incorporated with its representation before structure generation.

Table 1 shows that SoupFold improves structure prediction across all three base models, achieving state-of-the-art performance as illustrated in Figure 1. For protein-protein complexes, SoupFold consistently improves predictions. The largest gain is observed with ESMFold2, with a 6.93 percentage-point improvement at the high-quality threshold. SoupFold also improves protein-ligand prediction across all base models, with the largest gain for OpenDDE from 55.38 to 63.44. These results show that representations from other co-folding models provide useful information for improving both protein-protein and protein-ligand structure prediction. SoupFold also outperforms best-confidence selection over all 20 individual model's samples (All-Model-Best), avoiding the cross-model confidence misalignment inherent to sample-level ensembling.

We evaluate how teacher representations change prediction quality across protein-protein interfaces. Among the 274 protein-protein interfaces evaluated by DockQ, 41 move to a higher-quality category, 220 remain unchanged, and 13 move to a lower-quality category. We illustrate examples in Figure 3. Thus, representation transfer mostly preserves the original prediction while shifting more interfaces toward higher-quality categories than lower-quality ones. This allows SoupFold to recover base-model predictions without broadly disrupting already successful predictions.

2.2 Standalone Performance Does Not Always Determine Representation Utility

We study whether the benefits of a model as a source of representations are determined by its structure prediction performance. One might expect stronger co-folding models to always provide more useful representations. We find that this is not necessarily the case.

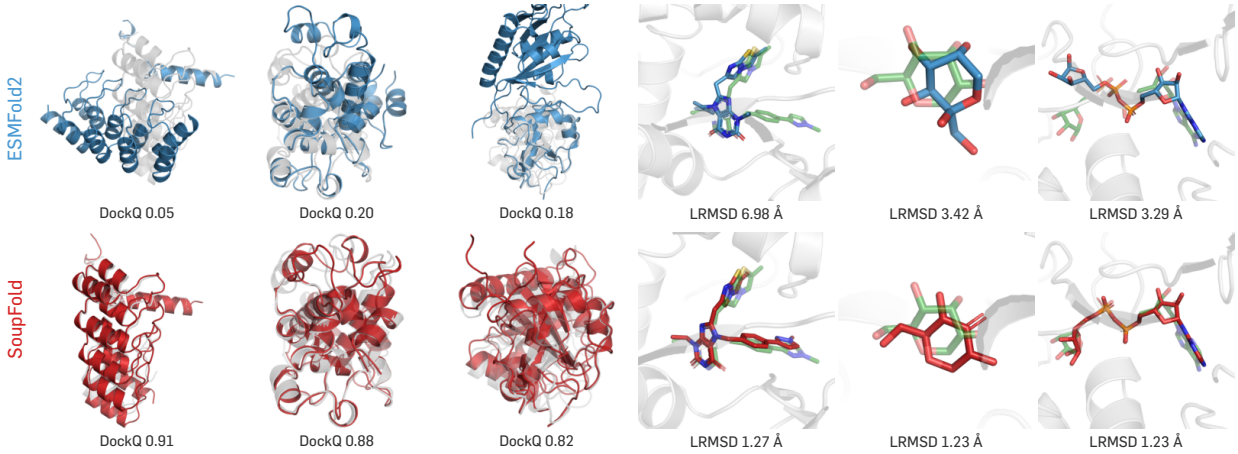

Figure 3 SoupFold recovers predictions using teacher representations. Examples of protein-protein interfaces whose prediction quality improves after representation transfer.

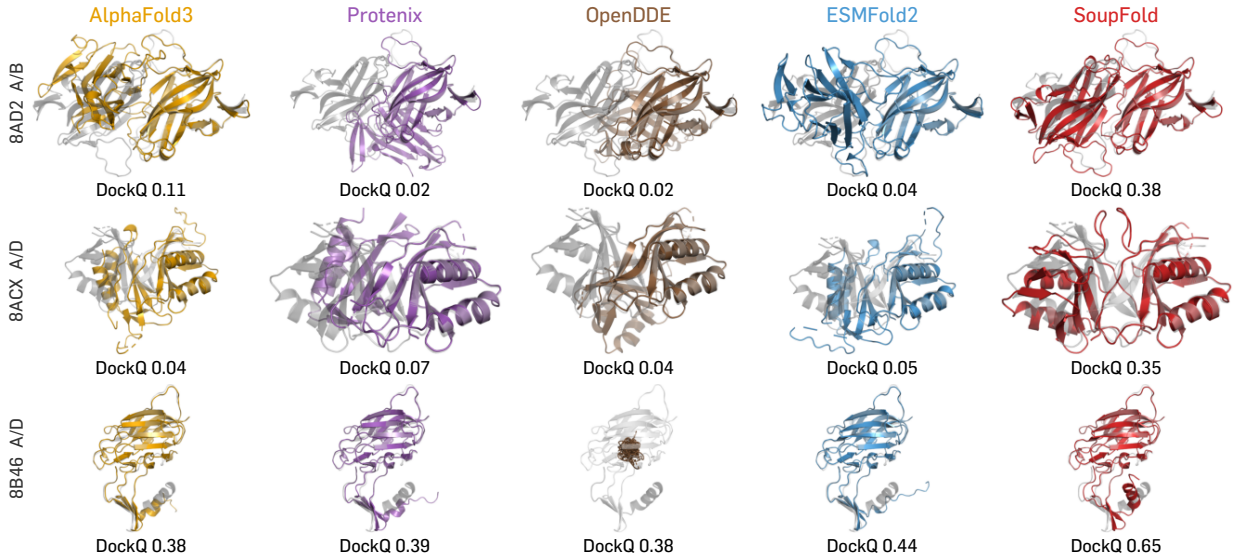

Figure 4 SoupFold succeeds even when all individual co-folding models fail. Examples of protein-protein interfaces for which all individual models fail to produce an acceptable structure, while SoupFold recovers acceptable predictions.

In Table 2, one can see that representations from models with lower standalone performance can still improve stronger models. This distinction between prediction quality and representation utility provides further evidence for complementarity across co-folding models. A model can encode structural patterns that are useful to another model even when it does not reliably convert that information into an accurate final structure. Combining models at the representation level can therefore exploit information that would be missed by comparing only their final predictions.

Table 2 Teacher ablation for SoupFold based on ESMFold2. P, A, and O denote Protenix, AF3, and OpenDDE, respectively.

Teachers	≥ 0.80	≥ 0.49	≥ 0.23
P + A + O	**46.35**	71.17	**77.37**
A + P	45.19	**71.85**	76.67
A + O	44.28	70.48	76.01
P + O	44.12	69.12	73.90

2.3 SoupFold Recovers Predictions Where Individual Models Fail

More interestingly, we also find cases where all individual co-folding models fail, but SoupFold successfully predicts the structure. We qualitatively illustrate these cases in Figure 4. Here, the improvement cannot be explained by transferring the representation of a teacher that already predicts the correct structure. Instead, combining representations from multiple models can recover useful information that is not sufficient for

successful prediction in any individual model. These results provide stronger evidence that co-folding models encode complementary information in their representations to output acceptable structure predictions.

3 Method

Given a set of co-folding models, we choose one model b as the base model, whose diffusion module produces the final structure, and use the remaining models $\mathcal{T}$ as teachers transferring information. For each teacher $t \in \mathcal{T}$, we learn a transfer network $f_{t\to b}$ from the teacher pair representation to the base pair representation using mean-squared error. Each transfer network $f_{t\to b}$ is a two-layer MLP with a hidden dimension of 1024. We present R^2 scores of each transfer network in Section A.1.

At inference time, we combine the representations by a weighted average in representation space. Each teacher representation $H^{(t)}$ is mapped into the base model's representation space through a teacher-to-base map $f_{t\to b}$, and the teacher representations are averaged with the base representation:

$$H = \frac{w_b H^{(b)} + \sum_t w_t f_{t\to b}(H^{(t)})}{w_b + \sum_t w_t}$$

where w_b and w_t are coefficients of the base model and each teacher. Then, the updated representation H is passed to the diffusion module to predict the structure. In all experiments, we use $w_b = w_t = 1$.

4 Related Works

Co-folding Models. Modern co-folding models predict the joint structure of biomolecular complexes containing proteins, nucleic acids, and small molecules. AlphaFold3 (Abramson et al., 2024) used a Pairformer trunk to construct pairwise representations of the complex, followed by a diffusion module that generates its atomic structure. This architecture has since been adopted by several open-weight co-folding models, including Boltz (Passaro et al., 2025; Wohlwend et al., 2024), Chai (Chai Discovery et al., 2024), Protenix (ByteDance AML AI4Science Team et al., 2025), OpenFold3 (The OpenFold3 Team, 2026), ESMFold2(Candido et al., 2026), and OpenDDE (Aureka AI OpenDDE project, 2026).

Representations in Co-folding Models. Several studies have studied internal representations of co-folding models. AlphaLink (Stahl et al., 2023) embeds crosslink-derived distance restraints and adds them to the pair representation of AlphaFold2, showing that modifying this representation can guide structure prediction toward conformations. More recently, Jang et al. (2026) repurpose representations from Boltz as atom-level molecular representations and show that they transfer to downstream tasks. Suzuki and Amagasa (2026) show that scaling the pair representations of AlphaFold3 changes their conformational sampling. In contrast to prior works, we transfer representations across multiple models.

Representation Distillation and Ensembling. Representation distillation transfers information by aligning intermediate representations between pre-trained teacher and student models (Ahn et al., 2019). Recent work has applied this idea to generative models for images (Yu et al., 2025) and protein structures (Kim et al., 2026). Beyond distillation, pretrained models can be combined by averaging their parameters, as in Model Soups (Wortsman et al., 2022), or by mapping representations across models (Bansal et al., 2021).

5 Conclusion

We studied whether independently trained co-folding models encode complementary information in their internal representations, and whether this information can be transferred across models to improve structure prediction. We introduced SoupFold, which combines pair representations from multiple co-folding models at inference time. SoupFold consistently improves structure prediction and can recover successful predictions even for targets where the individual models fail.

Our results also suggest several directions for future work. First, while our study focuses on protein-ligand and protein-protein complexes, SoupFold may extend to other complex classes. As initial evidence, Figure 1 shows that SoupFold improves OpenDDE on 160 FoldBench antibody-antigen targets even with representations from co-folding models that are achieved relatively lower performance on antibody-antigen prediction (consistent with observations of Section 2.2). Second, SoupFold currently combines teacher information using a fixed weighting scheme. Using target-dependent teacher weights may better improve the performance. Next, the transfer maps provide a simple way to align representations across models, but

more expressive mappings or representation distillation may further improve transfer. Finally, extending representation transfer to other biomolecular models, such as protein or molecular foundation models (Hayes et al., 2025; Kim et al., 2026; Méndez-Lucio et al., 2024), may provide additional information that is not captured by co-folding training.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9155–9163, 2019. doi: 10.1109/CVPR.2019.00938.

Aureka AI OpenDDE project. Folding, reasoning, and scaling with open-source drug discovery engine, 2026.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ak06J5jNR4.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - advancing structure prediction through a comprehensive AlphaFold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://www.biorxiv.org/content/early/2026/06/04/2026.06.03.729735.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Hyosoon Jang, Hyunjin Seo, Honghui Kim, Seonghyun Park, Taewon Kim, Yunhui Jang, and Sungsoo Ahn. A systematic evaluation of co-folding model representations for small-molecule learning, 2026.

Taewon Kim, Hyosoon Jang, Hyunjin Seo, Seonghwan Seo, Hyeongwoo Kim, Wonho Zhung, Mingyeong Shin, Wooyoun Kim, and Sungsoo Ahn. Atom-level protein representation learning improves protein structure prediction, 2026.

Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology*, 41(12):1810–1819, 2023. doi: 10.1038/s41587-023-01704-z.

Shosuke Suzuki and Toshiyuki Amagasa. Biasing conformational sampling in AlphaFold 3 and Boltz-2 via pair representation scaling. *Journal of Chemical Information and Modeling*, 2026. doi: 10.1021/acs.jcim.6c02094.

The OpenFold3 Team. OpenFold3-preview2 technical report, 2026. URL https://portal.openfold.omsf.io/reports/of3p2_technical_report.pdf. Technical report.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing

biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. Pmlr, 2022.

Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with FoldBench. *Nature Communications*, 17(1):442, 2026. doi: 10.1038/s41467-025-67127-3.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

A Additional Details

Figure 5 R^2 **scores of transfer networks**.

A.1 Training Transfer Networks

We train a transfer network $f_{t\to b}$ for each directed pair of ESMFold2, AlphaFold3, Protenix, and OpenDDE. Each map is a two-hidden-layer MLP of width 1024, applied independently to each pair entry. The corresponding pair dimensions are 256, 128, 128, and 384. Pair representations are extracted after ten recycling iterations and standardized per channel. Each map is trained with

$$\mathcal{L}(f_{t\to b}) = \mathbb{E}_{(\mathbf{z}^t,\mathbf{z}^b)\in\mathcal{D}} \left\| \mathbb{E}_{(i,j)} \left\| f_{t\to b}(\hat{z}^t_{ij}) - \hat{z}^b_{ij} \right\|_2^2 \right\|.$$

Here, $z^m_{ij} \in \mathbb{R}^{c_m}$ denotes the pair representation of model m for token pair (i, j), and $\hat{z}^m_{ij} = (z^m_{ij} - \mu^m)/\sigma^m$ is its standardized form. As part of the inference procedure, we first fit the transfer networks on the pair representations of the target complexes ($(\mathbf{z}^t, \mathbf{z}^b) \in \mathcal{D}$) before structure generation. Because the models use different tokenizations, we align pair entries using their token maps and retain only targets with exact correspondence across all four models. We use random crops of 512 tokens for 20 epochs with AdamW, a learning rate of 3×10^{-4}, no weight decay, and gradient clipping at 1.0. The same procedure can be applied to any collection of user-provided tasks. Figure 5 reports the R^2 of each directed mapping. An alternative direction can be pretraing the transfer networks on a held-out complexes and reuse them for unseen targets, removing the need to fit the mappings for each target set.